Thermodynamics of RNA structures by Wang-Landau sampling

Feng LOU & Peter CLOTE

ARENA, Boussens

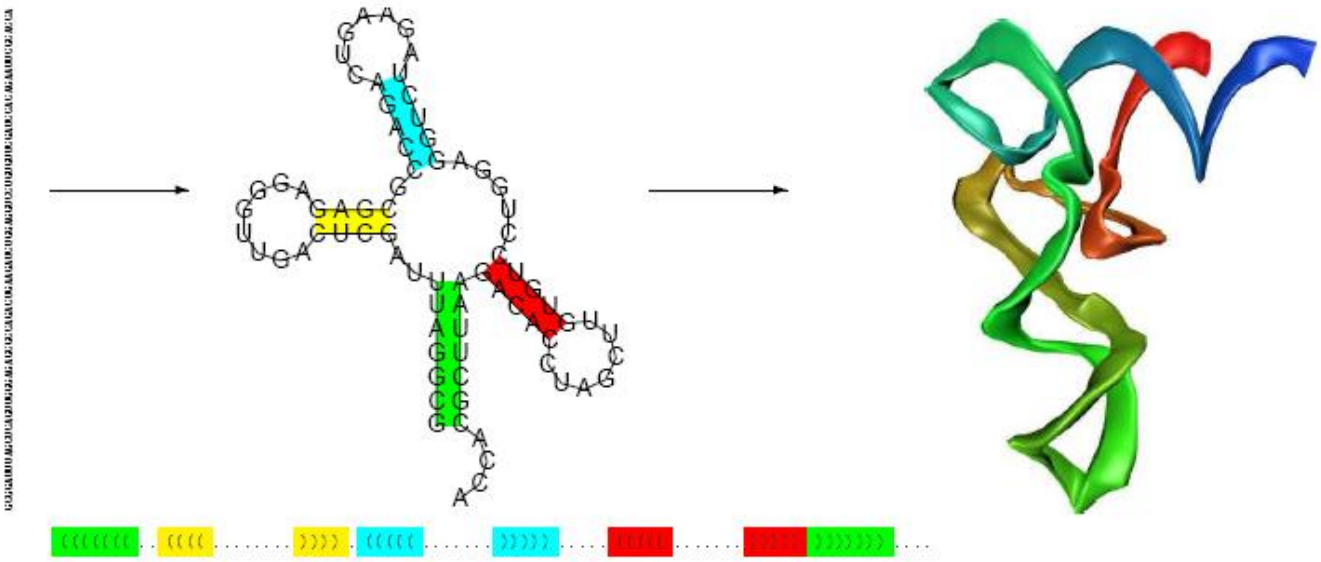
25/02/2010

Outline

RNA is an important biomolecule, now known to play both an information carrying role, as well as a catalytic role.

- The genomic information of retroviruses, such as the hepatitis C and human immunodeficiency viruses, is encoded by RNA rather than DNA.

- The peptidyl transferase reaction, arguably the most to whom correspondence should be addressed important enzymatic reaction responsible for life, is catalyzed not by a protein, but rather by RNA.



It is computationally intractable to compute the minimum free energy tertiary structure of RNA; indeed, determining the optimal pseudoknotted structure without any constraints on the type of pseudoknots is NP-complete.

In contrast, by disallowing pseudoknots, secondary structure prediction is algorithmically tractable; there are dynamic programming algorithms to compute the minimum free energy structure for a single RNA molecule, as well as for the hybridization of two RNA molecules, and even more than two RNA molecules.

Such thermodynamics-based algorithms are particularly important, since the tertiary structure of RNA is known to be largely determined by secondary structure, which acts as a scaffold for tertiary contacts;

METHODS: Boltzmann distribution

Boltzmann distribution weights each structure S of RNA ω by a Boltzmann factor: $B_{S,\omega} = e^{\frac{-E_{S,\omega}}{RT}}$ where :

- $E_{S,\omega}$ is the free energy of S (kCal/mol)
- T is the temperature in Kelvin
- R is the Gas constant $(1,986 * 10^{-3} kCal * K^{-1} * mol^{-1})$

The partition function of Boltzmann is calculated by:

$$Z_\omega = \Sigma_{S \in S_\omega} e^{\frac{-E_{S,\omega}}{RT}}$$

where S_ω is the ensemble structures compatibles with RNA ω

The Boltzmann probability of an structure S:

$$P_{S,\omega} = \frac{e^{\frac{-E_{S,\omega}}{RT}}}{Z_{S,\omega}}$$

METHODS: Monte Carlo algorithm

Generally, any (probability) distribution can be sampled by a **Monte Carlo** type algorithm.

Prerequisites: **Detailed Balance**

$$\pi(x)p(x \to y) = \pi(y)p(y \to x)$$

Metropolis rule: $p(x \to y) = min(1, \frac{\pi(x)}{\pi(y)})$

Boltzmann sampling: $\pi(x) = \frac{1}{Z} e^{\frac{-E_X}{RT}}$

Algorithm 1 Pseudocode for Metropolis-Hastings algorithm with simulated annealing

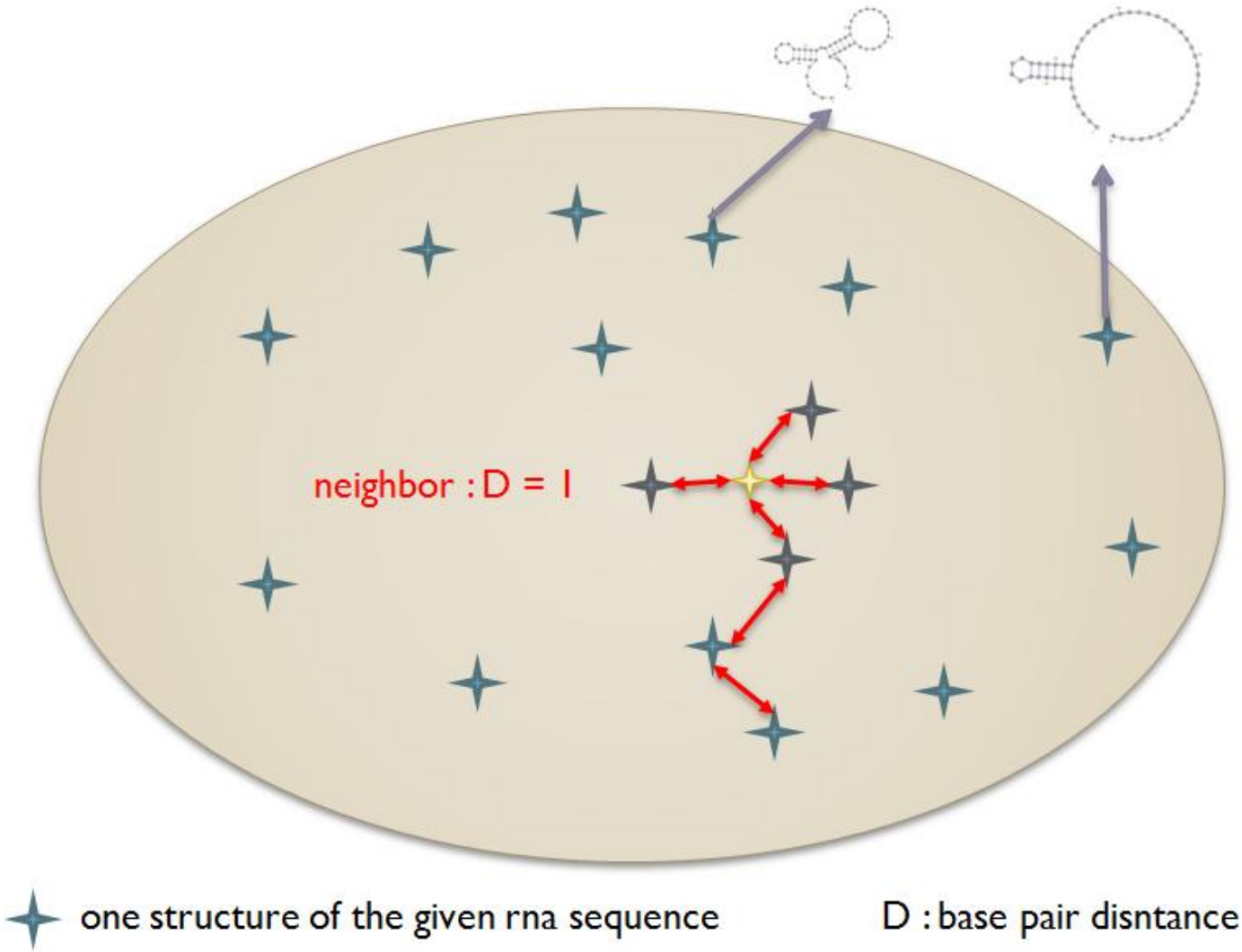
```
1: procedure MetropolisHastings
2:     T = T_hi
3:     x = initial state
4:     while T > T_low do
5:         repeat M times
6:             choose random neighbor y ∈ N_x
7:             if E(x) ≤ E(y) then
8:                 x = y
9:             else
10:                choose random z ∈ (0, 1)
11:                if z < e^(−(E(y)−E(x))/T) then
12:                    x = y
13:                end if
14:            end if
15:        until
16:        T = T * 0,9
17:    end while
18:    Return x
19: end procedure
```

METHODS: Wang-Landau sampling

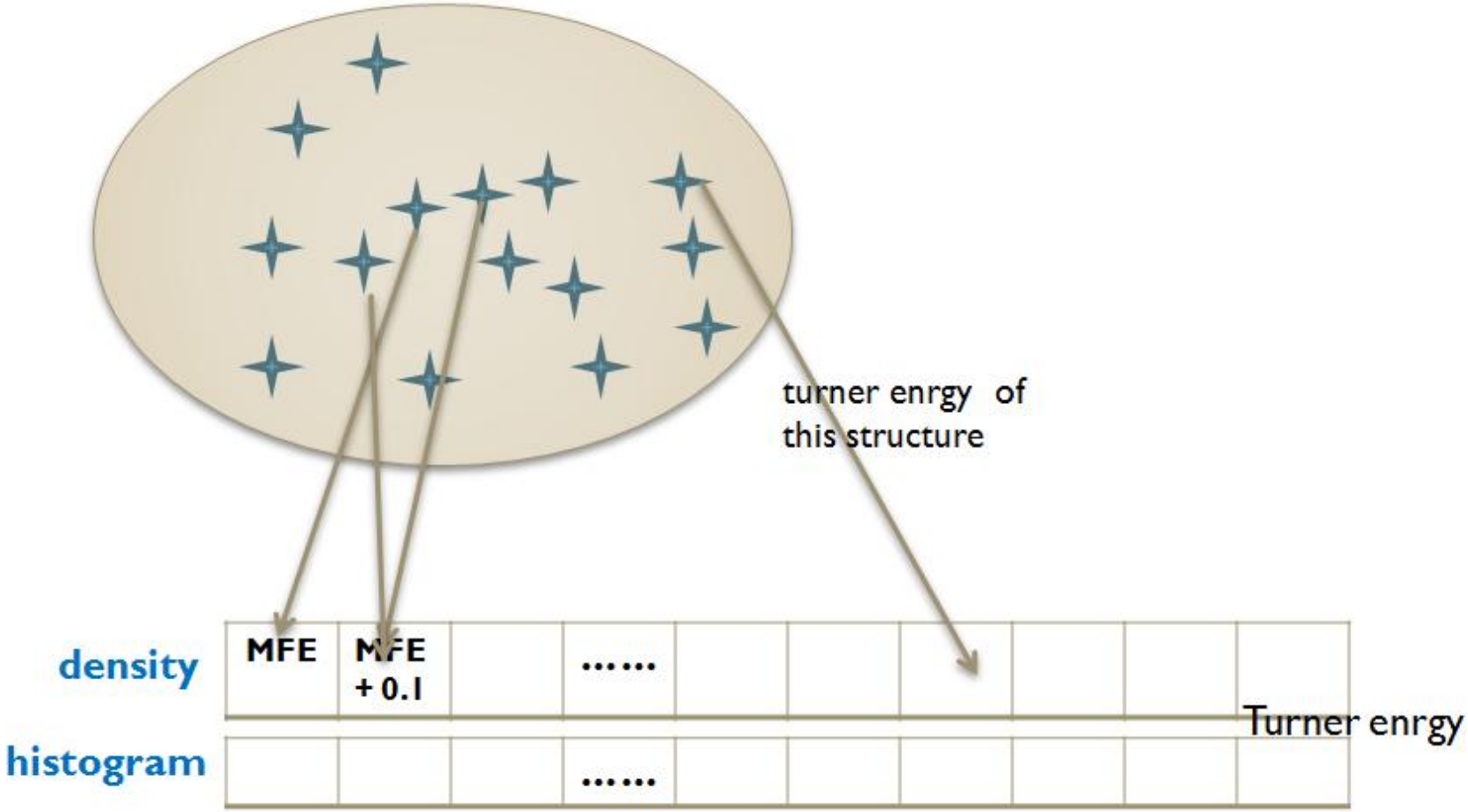
F. Wang and D. P. Landau.
Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. Phys. Rev. Lett., 86:20502053, 2001.

"A dynamic Monte Carlo algorithm to estimate the density of states by performing a random walk in energy space with a flat histogram"

Illustration of Wang-Landau random walk



How to get the Density Of State (d.o.s)



Starting from initial state (x), a random neighbor (y) is chosen with a transition probability:

$$P(x \rightarrow y) = min(1, \frac{g(E_x)}{g(E_y)})$$

g(Ei) : the density of state of energy Ei

- If the move is accepted, the value of $g(E_y)$ is multiplied with a modification factor c > 1 and the histogram entry $h(E_y)$ is updated.
- If the move is rejected, $g(E_x)$ is multiplied with c and $h(E_x)$ is Incremented.

Algorithm 2 Pseudocode for Wang-Landau algorithm, as applied to RNA secondary structure density of states computation.

```
1: procedure WANG-LANDAU
2:     S = φ // empty initial structure
3:     c = e // initial modification factor
4:     while c > 1 + ε do
5:         for all energy bins e: g(e) = 1
6:         while h is not flat do // h is flat, if for all x in h; we have x ≥ k * mean(h)
7:             repeatM times
8:                 choose random neighbor T ∈ NS
9:                 e_0 = bin(E(S)); e_1 = bin(E(T));
10:                choose random z ∈ (0, 1)
11:                if z < g(e_0)/g(e_1) then
12:                    S = T; e = e_1
13:                else// S remains unchanged
14:                    e = e_0
15:                end if
16:                g(e) = c * g(e) // update d.o.s
17:                h(e) = h(e) + 1 // update histogram
18:            until
19:        end while
20:        c = c^(1/2) // reduce modification factor
21:    end while
22:    return relative density of state: g
23: end procedure
```

RESULTS: Density of state for a single RNA energy

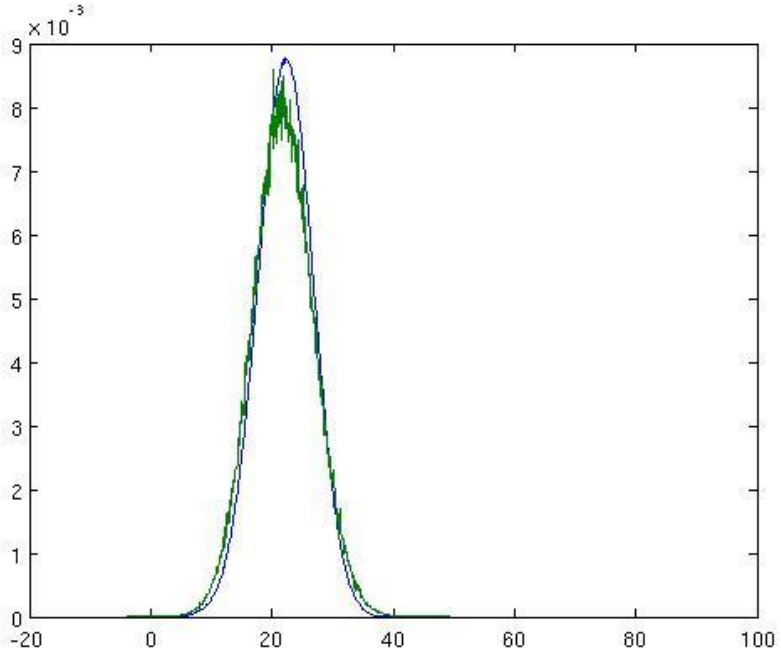


Figure: Validation Wang-Landau sampling by RNASubopt data

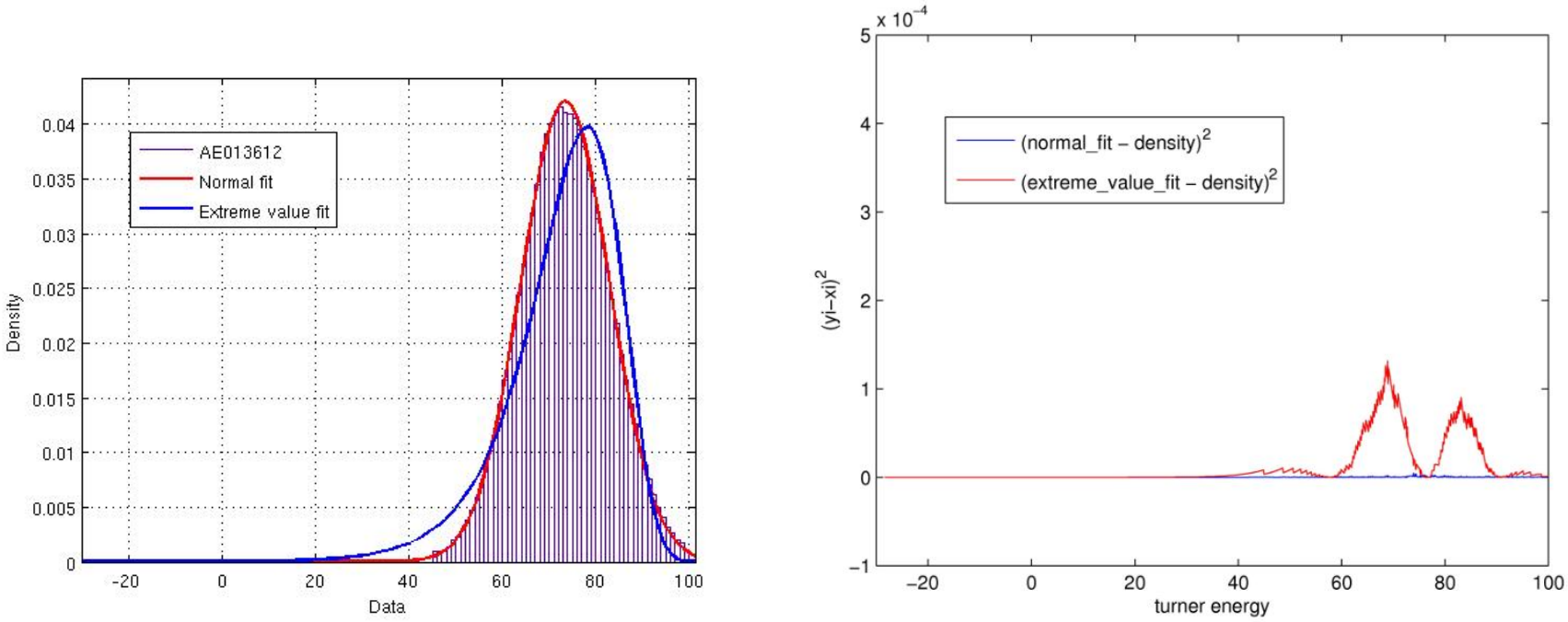


Figure: (Left)Density of states for free energy of secondary. (Right)Sum of squared differences between the density of states and the best fitting normal distribution resp. extreme value distribution.

Turner Energy	Relative Frequency	Secondary Structure
-3.3	2.363748736e-07	((.(((((....))))))).....
-3.0	2.667637567e-08	((.(((((....)))))))......(((.....)))
-2.8	2.3869453e-07	...(((((....))))).....
-2.5	5.121959212e-08	...(((((....)))))........(((.....)))
-2.2	1.275146796e-07	...(((((....)))))..(((.......))).....
-2.1	5.770032133e-08	..........((((..(((........)))..))))
-1.9	3.674825963e-08	((.(((((....)))))))........((.....)).
-1.7	1.333734872e-07	((.(((((....)))))))......(....).....
-1.6	2.363748736e-07	......(((((..............)))))....
-1.5	4.840633891e-07	((.((((......))))))..............
-1.4	7.658995002e-08	...(((((....)))))..........((.....)).
-1.2	2.135469025e-07	((.((((......)))))).....(((.....)))
-1.1	3.294584489e-07	...((((........)))).....
-1.0	3.107089098e-07	...((((......)))).....
-0.9	3.548407128e-07	((.(((((....)))))).......)).....
-0.8	6.239823716e-07	...(((((....)))))..((.........)).....
-0.7	5.082099809e-08	...((((......))))........(((.....)))
-0.5	4.330649611e-07	((.((((........))))).....)).....
-0.4	1.453421863e-07	...((((......)))).(((.......))).....
-0.3	3.569259607e-07	...(((((....)))))..........(.....)..
-0.2	5.233191908e-08	.(.(((((....)))))...)....(((.....)))

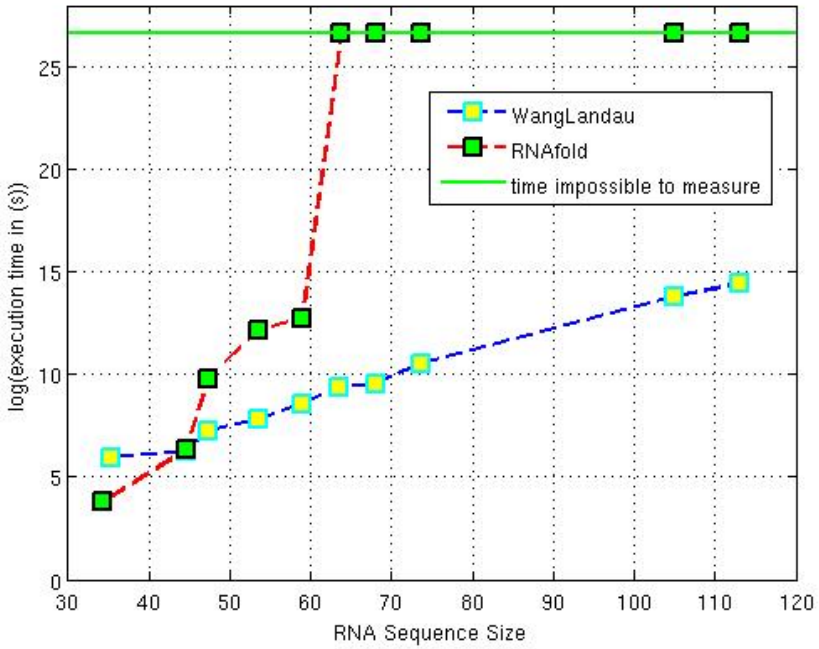


Figure: (Left)Density of states for free energy of secondary. (Right)Sum of squared differences between the density of states and the best fitting normal distribution resp. extreme value distribution.

RNAsubopt: calculate suboptimal secondary structures of RNA including in Vienna Package. with option -e -D.

Boltzmann Partition function:

$$Z_\omega = \Sigma_{S \in S_\omega} e^{\frac{-E_{S,\omega}}{RT}} \approx \Sigma_E g(E) * e^{\frac{-E}{R*T}}$$

Gibbs free energy:

$$G = -R * T * \ln(Z)$$

$$G = H - T * S$$

RESULTS: Melting temperature for hybridization of two RNA:

Melting temperature:

The temperature at which one half of the strands are unhybridized and unfolded.

Predicting the melting temperature of RNA duplexes is one of the most important applications of the partition function for interacting nucleic acid pairs.

Melting experiments have been the most useful way to measure the stabilities of RNA and DNA structures under different conditions.

The number of secondary structures:

- RNA sequence of length n
- $\delta_{i,j} = 1$ if positions i, j can form a Watson-Crick or wobble pair, otherwise let $\delta_{i,j} = 0$
- $\theta = 3$ denote the minimum number of unpaired bases in a hairpin loop.
- $N_{i,j}$ denotes the number of secondary structures on subsequence [i,j] of the given RNA sequence.

We have that:

- IF $j < i + 3, N_{i,j} = 1$
- ELSE:

$$N_{i,j} = N_{i,j-1} + \Sigma_{k=i}^{j-\theta-1} \delta_{k,j} * N_{i,k-1} * N_{k+1,j-1}$$

$\delta_{k,j} = 1$, if nucleotide k and j are base pair, else $= 0$.

The total number of secondary structures is then $N_{1,n}$.

This algorithm is $O(n^3)$

The number of hybridizations:

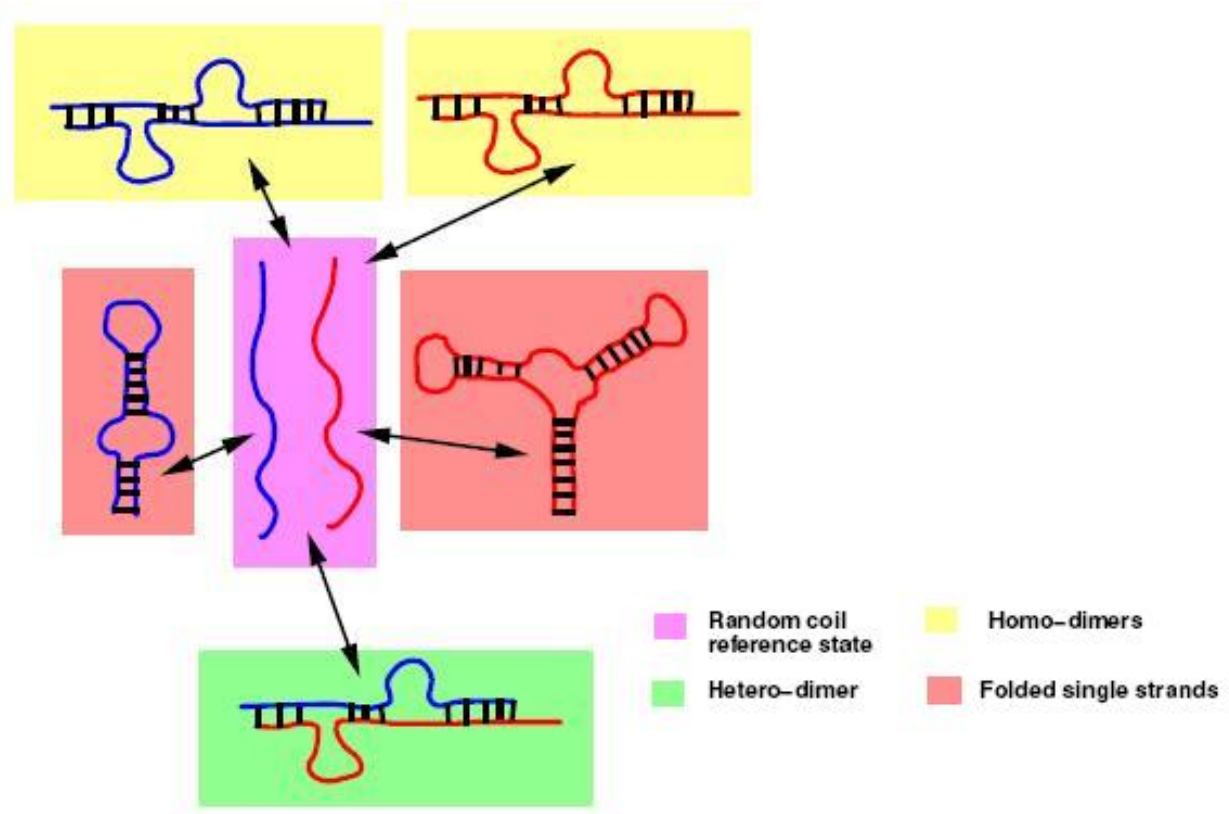


Figure: The ensemble of seven possible species: unfolded A and B, folded A and B,and hybridized A-A, B-B and A-B is depicted here.

- an RNA sequence A = $a_1, \ldots, a_n$ of length n.
- an RNA sequence B = $b_1, \ldots, b_m$ of length m.
- let $HP_{i,j} = 1$ if positions a_i, b_j can hybridize, forming a Watson-Crick or wobble pair, otherwise let $HP_{i,j} = 0$.
- For 1 = i, j = n, 1 = k, l = m, let $H_{i,j;k,l}$ denote the number of Hybridizations of the subsequence $a_i, ..., a_j with b_k, ..., b_l$

If ($j < i$ or $l < k$), then define $H_{i,j;k,l} = 0$.

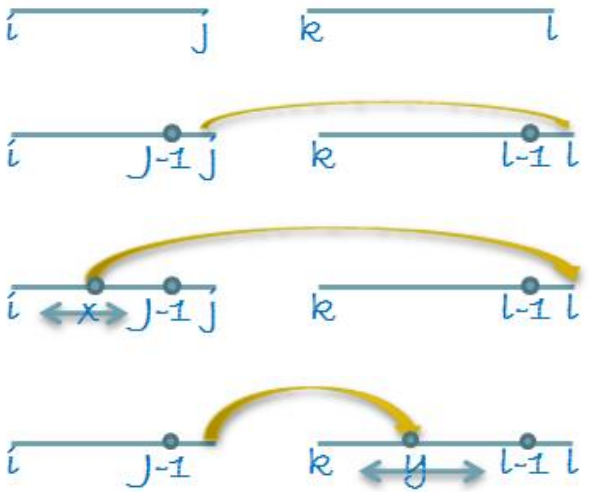
Otherwise, define $H_{i,j;k,l}$ by:

$$H_{i,j;k,l} =$$

$$H_{i,j-1;k,l-1} * (1 + HP(j,l))$$

$$+ \sum_{x=i}^{j-1} HP(x,l) * H_{i,x-1;k,l-1} * NA_{x+1,j}$$

$$+ \sum_{y=k}^{l-1} HP(j,y) * H_{i,j-1;k,y-1} * NA_{y+1,j}$$



The total number of pseudoknot-free hybridizations is then $H_{1,n;1,m}$. This algorithm is $O(n^4)$

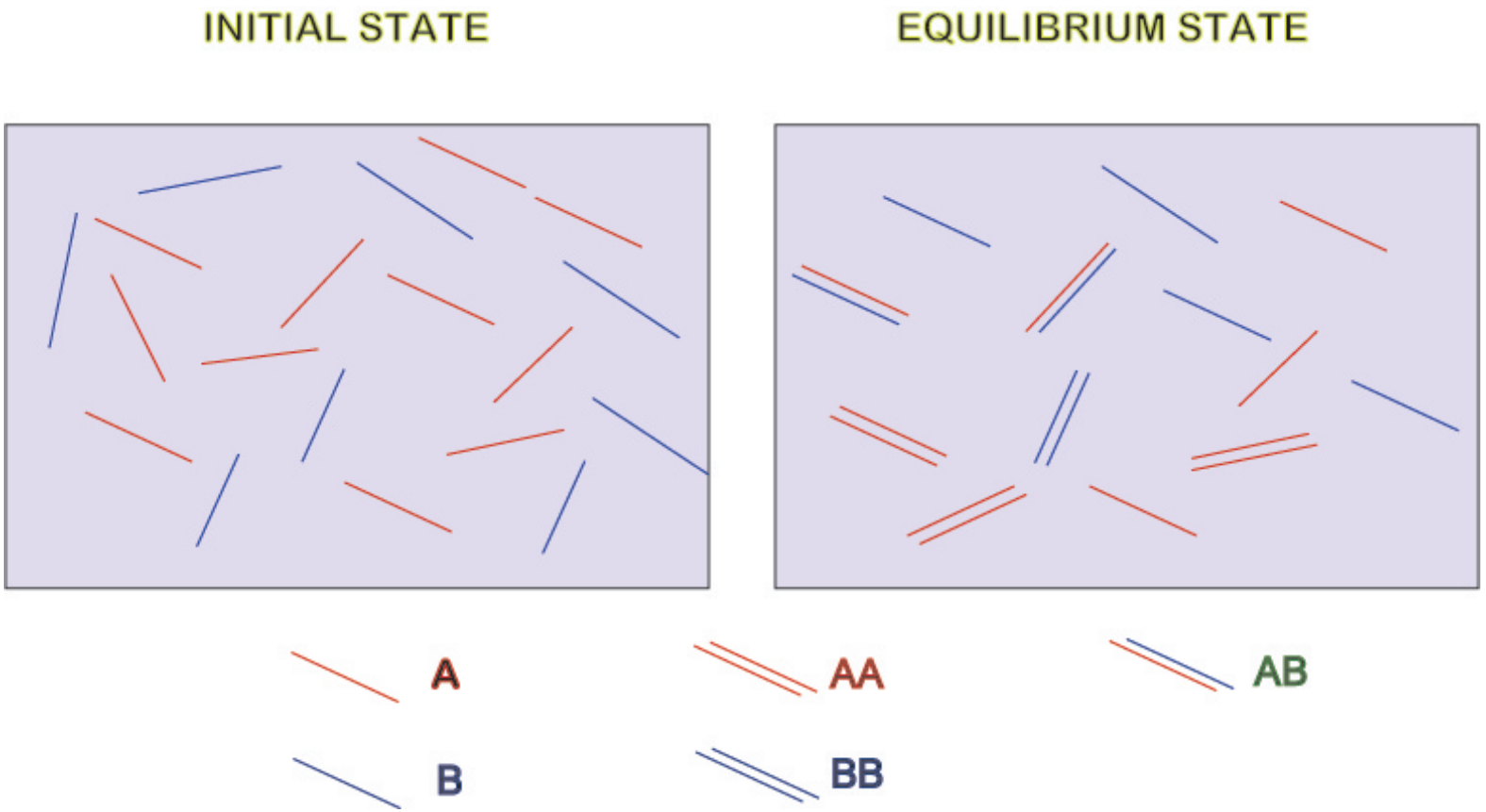


Figure: Illustration of hybridization of two RNA molecules.

Ensemble free energy:

1. Compute the temperature independent **number of structures** : N(A), N(B), and **number of hybridization** : N(AA), N(BB), N(AB) for each of the five species A, B, AA, BB, AB.

2. For each temperature T in Celsius from 0 °C to 100 °C, run program WL to compute the **relative density of states** f(A,T), f(B,T), f(AA,T), f(BB,T), f(AB,T) for each species.

3 From previous step (1) and (2), compute the **absolute density of states** and using following equation, compute the temperature dependent partition function Z(A, T), Z(B, T), Z(AA, T), Z(BB, T), Z(AB, T) for each species.

$$Z(T) = \Sigma_E g(E) * e^{\frac{-E}{R*T}}$$

where g(E) is the absolute density of state for energy E.

4 For each temperature T from 0 °C to 100 °C, compute the **ensemble free energy** ΔG(T) from the partition functions for each of five species. *(Dimitrov and Zuker 2004)*

a Redundancy correction:

$$Z_{aa} = Z_{aa} - Z_a^2$$
$$Z_{bb} = Z_{bb} - Z_b^2$$
$$Z_{ab} = Z_{ab} - Z_a * Z_b$$

b Symmetry correction:

$$Z_{aa} = \frac{Z_{aa}}{2}, \quad Z_{bb} = \frac{Z_{bb}}{2}$$

c Chemical equilibrium constants:

$$K_A = \frac{Z_{aa}}{Z_a^2}, K_B = \frac{Z_{bb}}{Z_b^2}$$

$$K_{AB} = \frac{Z_{ab}}{Z_b * Z_b}$$

d Concentration of molecules A and B:

$$2 * K_A * N_A^2 + K_{AB} * N_A * N_B + N_A - N_A^0 = 0$$

$$2 * K_B * N_B^2 + K_{AB} * N_A * N_B + N_B - N_B^0 = 0$$

e The ensemble free energy:

$$\mu_a = R * T * (\ln(\frac{N_A}{N_A^0} - \ln(Z_A)))$$

$$\mu_b = R * T * (\ln(\frac{N_B}{N_B^0} - \ln(Z_B)))$$

$$F = \mu_a * N_A + \mu_b * N_B + \mu_{aa} * N_{AA} + \mu_{bb} * N_{BB} + \mu_{ab} * N_{AB}$$

which can be simplified to :

$$F = \mu_a * N_A^0 + \mu_b * N_B^0$$

f Normailization the ensemble free energy:

$$\Delta G = \frac{\mu_a * N_A^0 + \mu_b * N_B^0}{max(N_A^0, N_B^0)}$$

5 Calculate the **heat capacity** *(N.R markham Dissertation 2006)*

$$C_p = \frac{d\Delta H}{dT} = -T * \frac{d^2\Delta G}{dT^2}$$

The second derivative is computed numerically by fitting a parabola to 2m+1 evenly spaced points, using the approximation:

$$\frac{d^2\Delta G}{dT^2} \approx \frac{30}{m(m+1)4m^2(2m+3)\delta T^2} \sum_{-m \leq i \leq m} (3i^2 - m(m+1)\Delta G(T_0 + i\delta T)$$

The ensemble heat capacity is of interest because the local maximum (or, more generally, the local maxima) define the melting point(s) Tm(C_p).

6 In a post-processing step, smooth the heat capacity curve by computing a running average.

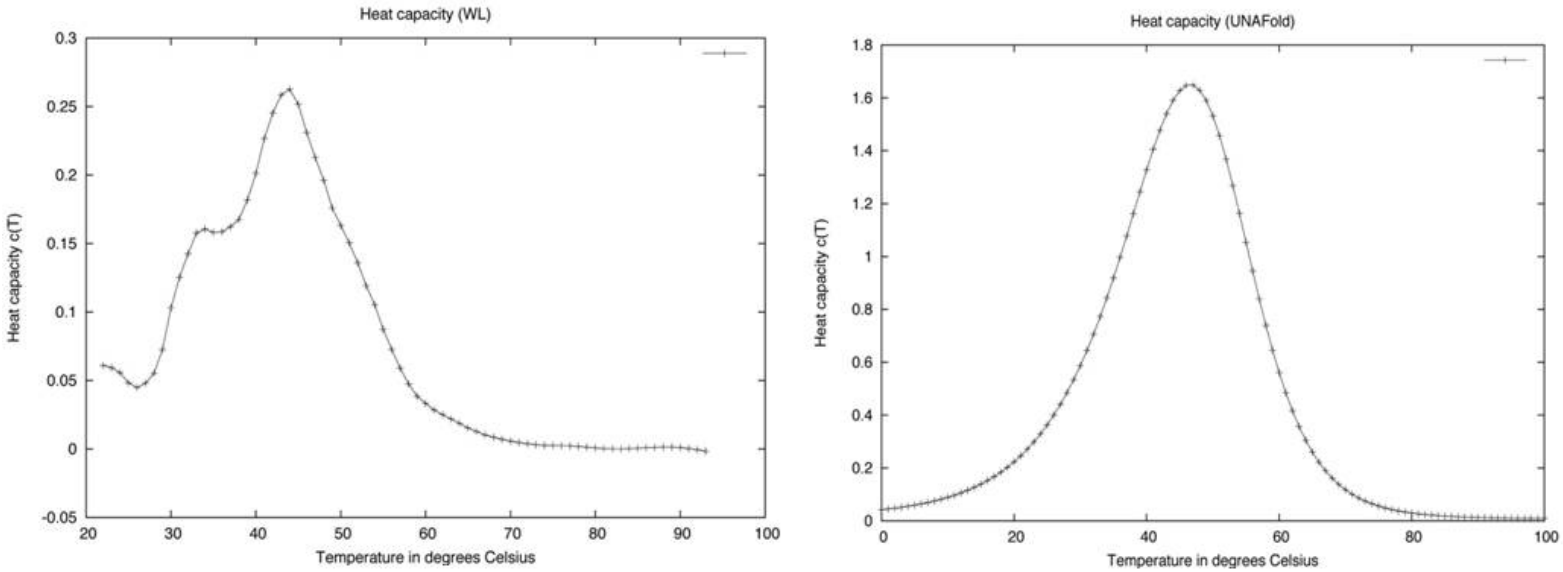


Figure: Heat capacity curve Cp(T) for the same sequence, computed by the WL(left) method described in this paper and Unafold(right).

Sequence	Experiment	Unafold	RNAcofold	WL
ACGCA & UGCGU	29.8	42.64	46.14	**42**
GCACG & CGUGC	37.5	46.61	43.91	44
AGCGA & UCGCU	30.2	42.88	45.15	**41**
GCUCG & CGAGC	37.2	47.75	44.71	48
ACUGUCA & UGACAGU	48.2	56.8	57.59	**51**
GUCACUG & CAUGUAC	51.1	58.44	55.91	56
AGUCUGA & UCAGACU	45.7	56.4	56.68	**52**
GACUCAG & CUGAGUC	52	59.11	56.25	**52**
GAGUGAG & CUCACUC	53.7	59.07	56.00	58

The date Experiment extract from article: Xia et al. (1998)
The date Unafold and RNAcofold extract from article: Chitsaz & al (2009)

DISCUSSIONS

- the advantage of WL over existent methods in computing the density of states for both single RNA molecules and for hybridization complexes of two RNA molecules.
- the program UNAFold does not allow any intramolecular structure (base pairing between two nucleotides of the same structure), a feature that our WL method permits, as does the RNAcofold program.

- the melting temperature TM computed by WL agrees reasonably well with that computed by $O(n^3)$ methods UNAFold, RNAcofold, and the recent $O(n^6)$ method of Chitsaz et al., each of which methods admits slightly different interactions.
- finally, we intend to implement a new energy evaluation function, that allows arbitrary pseudoknots, zig-zags, etc. This will allow us to estimate the partition function, ensemble free energy, heat capacity, melting temperature, etc. for a context known to be NP-complete.